Analysis of sea water infiltration in a sewage treatment plant using Random Forests and variable importance measures

Computational Mathematics Master's Final Project

Student: **Cristóbal Rodero Gómez**

Tutor: **Irene Epifanio López**

"Errors using inadequate data are much less than those using no data at all."

Charles Babbage, mathematician, philosopher, inventor and mechanical engineer.

"Correlation implies causation."

Said no statistician. Ever.

Abstract

Infiltration of salt water in water-treatment plants is a current challenge in Spanish coasts. This is a problem because when the sea water enters into the plant, it damages the filters and compromise the quality of the water filtration. In order to detect this water infiltration, the usual approach is to measure conductivity of the flows (ability to conduct electricity).

The work subsequently described in this document is a project funded by a research internship of the Universitat Jaume I's *Cátedra FACSA de Innovación del Ciclo Integral del Agua*. The goal of this project is to detect when a saltwater infiltration has occurred and to detect the most relevant variables which are related with the rising of water conductivity.
The approach chosen to deal with this problem has been the technique of Random Forests, a family of algorithm based on decision trees. The reasons of this elections are mainly the flexibility with respect missing data that Random Forests allow; the "black box"-like behaviour that does not need an *a priori* knowledge of the data structure; and the ability to explore the variables' importance through several measures.

In this project the mentioned methodology will be explained in detail, as well as the results obtained and the conclusions that follows them.

Keywords

Random Forest, Variable Importance Measure, Sea Water Infiltration, Water Conductivity Prediction

Acknowledgements

I would first like to thank Irene. There have been a lot of e-mails, threads and debates through this last year and each one of them has been indubitably useful (and not only in the academic sense). I am extremely thankful and indebted to her for her patience, for sharing their expertise and for their sincere and valuable guidance.

Obviously, I cannot forget about thanking Pau for his support over this insane year and with this project, for our mathematical debates and for the healthy competitiveness for learning more and more with each curiosity we bumped into. You will not *lose* any coming update.

Finally, I would be very hurt if I forgot expressing my very profound gratitude to my mother for providing me with unfailing support and continuous encouragement throughout my years of study.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lend me their hand in this venture.

This accomplishment would not have been possible without of all you.

Thank you.

Cristóbal

Contents

List of Figures

Chapter 1

Introduction and motivation

In this first chapter we will explain the context where this project has been developed, as well as the motivation that allow us to ensure that we are dealing with a relevant problem and finally the project's structure.

Infiltration of salt water in water-treatment plants is a current challenge in Spanish coasts. In some cases, these infiltrations have caused even faecal discharges [2]. This is because the sea water enters into the plant damaging the filters and compromising the quality of the water filtration.

The available literature is acknowledging the issues the municipalities are going to face regarding their waste-water management [3, 4]. Moreover, risks to coastal waste-water collection systems from sea-level rise have also been studied, although focusing on the effect of climate change [5].

In order to detect this water infiltration, the usual approach [6] is to measure the conductivity of the flows (ability to conduct electricity). The conductivity of sea water depends on the number of dissolved ions per volume (salinity) and the mobility of the ions (temperature and pressure). This measure has been highly recommended for more than ten years [7].

1.1 Objectives and methodology

The objective is, therefore, to implement a comprehensive program to identify current and future saltwater intrusion and so in the future we can take measures to stop these sources of intrusion. We have focused on identifying some of the most relevant variables that affect water conductivity.

The work subsequently described in this document is a project funded by a research internship of the Universitat Jaume I's (UJI) *Cátedra FACSA de Innovación del Ciclo Integral del Agua* [8].

Due to the aforementioned situation, FACSA has considered that a grant of introduction to research is necessary, in order to work jointly with a student of the Computational Mathematics'

Master.

The general goals to be achieved along this project are related with the problem of the sea water filtration to the water treatment plants. In order to study this topic, we will use the water's conductivity, since its raising could be interpreted as a sign of the presence of salts resulting from sea water infiltration.

Regarding the methodology, for the reasons we further expose, we have chosen to use Random Forests for regression. We decided to use this family of methods fundamentally for several reasons:

- Flexibility with respect to the form of the data. There is no need of having an internal pattern in the response (conductivity). Nevertheless, if it exists, Random Forests will benefit of it.
- It is well-behaved with missing values. We might have some missing values on the input variables due to miss-measurements or other glitches. This is not an issue with this technique.
- Possibility of exploring the relations between variables. This may help us find some explanations and, in future works, solve the water treatment plant problem.
- Previous work in the field of regression of temporal series with Random Forests has obtained good results. See for example [9].

1.2 Project's structure

This project is organised as follows: In chapter 2 we start giving some details about the company we have been working with, FACSA, and about the data used. Afterwards, in chapter 3 we will explain the methodology used for dealing with our problem, Random Forests, and some previous concepts needed. We will also explain here some of the key concepts: Variable important measures, in section 3.3.1. The results of the analysis done in this work are given in chapter 4. In the last chapter, chapter 5, we will end by giving some last remarks about what has been accomplished in this master's thesis and we will point out some possible lines of future work.

Chapter 2

About FACSA

In the previous chapter we motivated the problem, remarking the importance of sea water infiltration in water treatment plants. We now focus on one specific water treatment plant in a mediterranean coastal town. This plant is property of FACSA, a company with headquarters in Castellón.
Therefore, in this chapter we will give some details about the company we have been working with, describing their main purposes. We will also explain the data they provide us, remarking their reasons and the additional datasets we have used.

2.1 Company description

The company *FACSA* was founded in Castellón in 1873 with the aim of providing the province's capital city with a state-of-the-art water supply distribution network. Since then, they have expanded their activities and their presence has been consolidated all over Spain: Valencian Community, Aragón, Murcia, Castilla-La Mancha, La Rioja and Balearic Islands. They have become a model company in the water sector. Currently, they are providing water supply to more than a million people in seventy towns and villages every day [10].

FACSA offers all of the typical services of the water integral cycle: water gathering, purification, treatment, distribution and a further gathering and purification of sewage.

Quoting [10], FACSA's keys of success are betting for the continuous updating of their employees' knowledge and a team comprising staff of more than eight hundred multidisciplinary professionals in order to offer a high quality service focused on providing efficient answers and solutions.

One of the main FACSA's strategical goal is to work in a professional and constant way in order to fulfil a whole satisfaction with the client. Likewise, they promote, ease and stimulate the teamwork and collaboration between the company employees. Moreover, another objective is to offer a service enviromental-friendly. This is the reason why they work with interest groups who promotes a sustainable development [10].

The different departments FACSA accounts are shown in table 2.1.

FACSA'S DEPARTMENTS
Recruiting
Purchasing
Meter box verification
Accounting
Billing
Readers inspection
IT
Expansion and development
Innovation and ICT projects
Sewage treatment
Quality
Technical: • Maintenance • Delineation • Supply
Sales
HR
Management

Table 2.1: FACSA's departments.

FACSA belongs to a company group called *Grupo Gimeno*, divided in three sectors: *Gimeno Servicios* (services), *Gimeno Construcción* (construction) and *Gimeno Turismo y Ocio* (tourism and leisure). See figure 2.1 for more details.

As an outcome of the efforts and the great innovative ability, Grupo Gimeno possesses more than thirty companies operating all over the national territory with more than four thousand professionals taking part in their personnel.

The Grupo Gimeno's companies, besides operating in a national level, they also do it internationally through their presence in areas such as Southamerica and Saudi Arabia. They have headquarters in Castellón, Alicante, Valencia, Teruel, Murcia, Madrid, Barcelona, Mallorca, Zaragoza, Toledo, Sevilla and Pontevedra [11] (see figure 2.2).

2.2 Data acquisition

Returning to the main discussion, we need some quantification of the sea water infiltration. Since, to the best of our knowledge, there is no scientific agreement on a clear indicator of this fact, we followed the suggestions made by FACSA. They sent us three spreadsheets: one about electrical conductivity, another one about measures of water flow (discharge) that passes through the plant and a last one about daily rains.

Regarding the first two variables, as we aforesaid, the first one was the inflow (to the plant) conductivity. This is a purely analytic value, since they measure it in FACSA's laboratories using

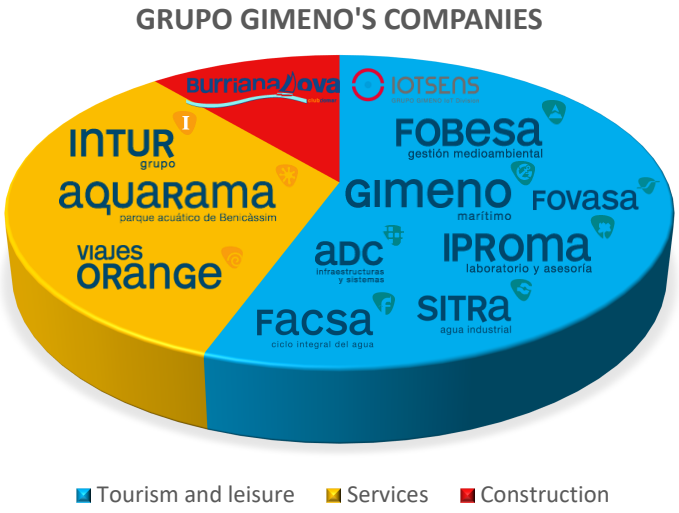

Figure 2.1: Schematics of the companies that belong to Grupo Gimeno. The divisions are made according to the work sector.

electrical conductivity meters. In this way, these values give us a quantification of the sewage ability to conduct an electric current. Sewage with salts, bases and acids may have higher conductivity coefficients than sewage with organic compounds that do not dissociate, where is nearly null. Nevertheless, this latter case is the one they deal with the most, although for our case, FACSA has considered this variable to be relevant for studying it and monitoring it. Their hypothesis is that, due to sea water infiltration, the salt concentration in water raises enough to characterise this fact by means of the conductivity coefficients.
Regarding the other variable, it reflects the amount of water volume that goes through the different processing units of the facility, finally treated in the waste-water treatment plant.

Both variables are measured at the entrance of the plant. Previously, the water is pre-treated in order to remove the bigger particles, the sand and the grease. Neither of the two stages are outdoors. The sieves are located inside a building and the sand removal and degreasing machines are covered by a four meters high dome, in order to allow access to the zone, but it confines the possible emission of gases.
Nevertheless, some water evaporation might exist, since in the sieves there are some waterfalls and the canals inside the dome are open-wide. This fact, together with aspiration systems for deodorising the emanated gases, should be taking into account for a further analysis. In this project we will assume this evaporation to be negligible.

These two datasets actually belong to *EDAR Benicàssim*. EDAR is a facility in Benicàssim whose initials in Spanish mean waste-water treatment plants. With a mean discharge value of 6715 m^3/day they serve more than six thousand people (see [1]). They make water to pass through different treatment processes, separated in two lines. In the water line they do:

- Pretreatment:

Figure 2.2: Approximate location of the different delegations of the Grupo Gimeno's companies.

  - Sieving.
  - Sand removal.
  - Degreasing.
- Primary treatment:
  - Physico-chemical.
  - Decantation.
- Secondary treatment:
  - Activated sludge.
  - Phosphorus elimination.

And in the sludge line:

- Thickener:
  - Gravity.
- Stabilisation:
  - Aerobic.

  - Stabilisation with lime.
- Dehydration:
  - Filter.

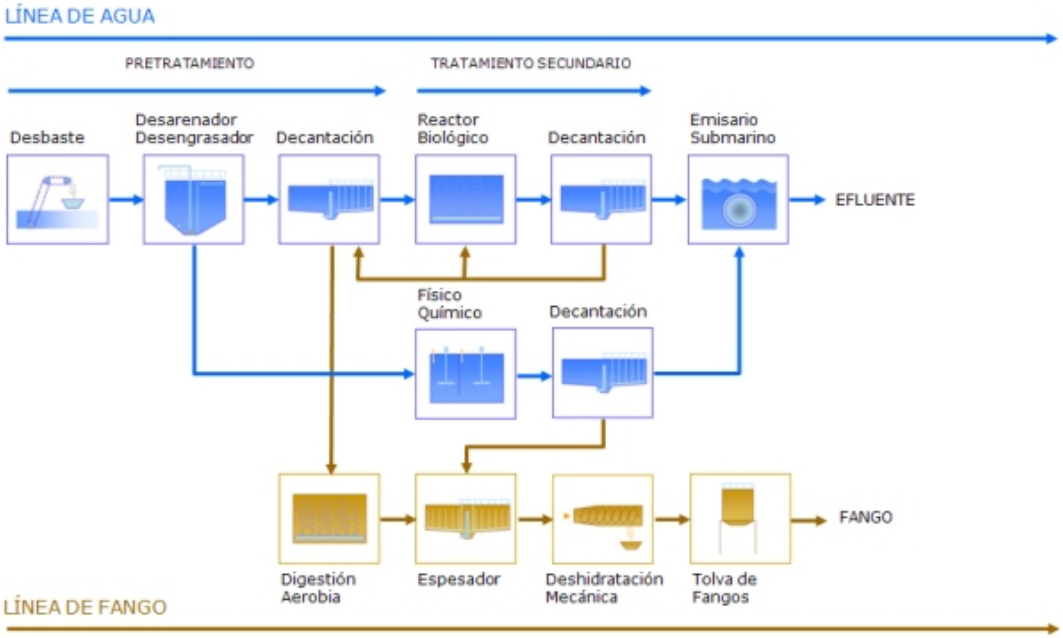

Figure 2.3: EDAR's processes block diagram. In blue, the water line; in yellow, the sludge line. Figure extracted from [1].

Although we have enumerated these processes, we will not detail them since it is out of the scope of this work. In figure 2.3 we can schematically see the water route through the different processes.

Now, the idea is consequently trying to predict the water conductivity using some independent variables, some *predictors*. The most direct approach could be to use the conductivity historic records (of the past six years) in order to try to model some intern pattern followed by the data. But, *a priori*, the data may not have any structure, so other variables should be studied as well. Exploiting the EDAR's datasets, we could use the flow rate likewise.

But we decided to test the hypothesis suggested by FACSA: due to intense rains, heavy streams of water drag sea water to the sewage treatment plant, rising therefore the conductivity. We have had three possible sources for pluvial data:

1. A FACSA's intern record. They facilitate us another spread-sheet with an integer value corresponding to the total rain's measurement for each day.
2. The State Meteorology Agency (AEMET). This agency has a service called *AEMET Open-Data* (see [12]) created for the diffusion and reuse of AEMET's information. The issue with

this approach was that, after obtaining a key, we found two main problems:

- There was, again, a single float value for amount of rain every day. We could buy the hourly data, so we were not interested in this source.
- Moreover, the available data was only for the last thirty days but we needed more previous data.

3. The observatory of the Castellon's planetarium (see [13]). Located at approximately six kilometres from FACSA headquarters (see figure 2.4), they provide values for data such as temperature, rain, wind or pressure at intervals of ten minutes. Although there were some missing values (at certain times, or certain days) these were distinguishable and this was not very common to occur.

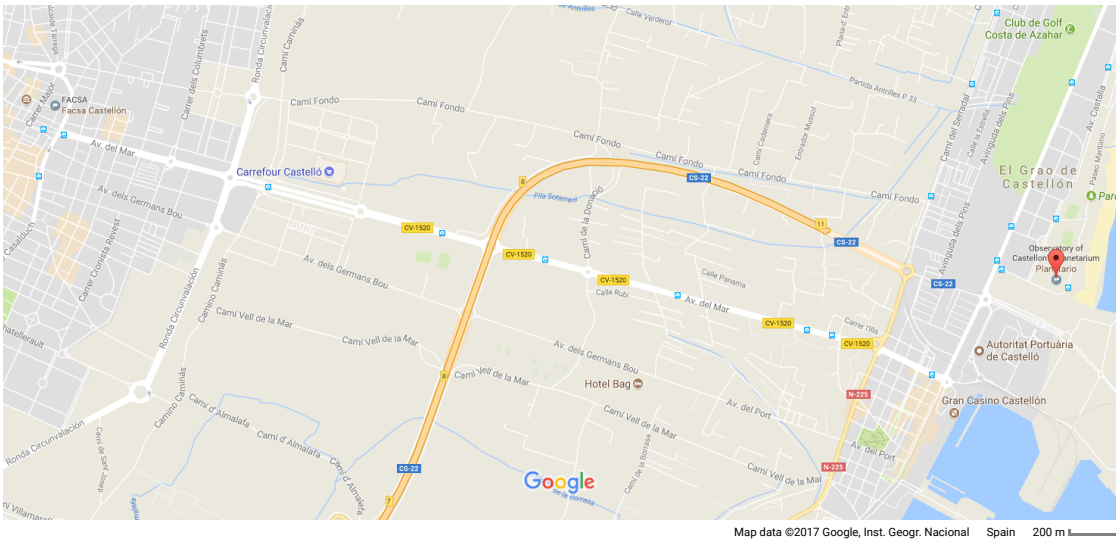

Figure 2.4: Map showing the location and distance between FACSA's headquarters (on the left) and the observatory of the Castellon's planetarium (on the right, pinned).

With these approximations, we chose the third one for its availability and usefulness. Furthermore, thinking long-term, this data structure could be advantageous for an approach with functional data.

As a final step of the pre-processing stage we removed the days where there was no data for conductivity.

The extracted values (for every analysed day) have been:

- Accumulated rain: for each day, we add up all the ten-minutes rain measures.
- Rainfall intensity: we took the maximum value of rain measure and divide it by ten. The goal is to quantify the torrential nature of rainfall.

- Average temperature of the day.

Now we have a clear knowledge of the data used, in the next chapter we will explain the methodology used for dealing with this variables, Random Forests, and some previous concepts needed.

Chapter 3

Methodology

In this chapter we will explain the main technique used for solving the problem of sea water infiltration: Random Forests. We decided to use this family of methods fundamentally for several reasons:

- Flexibility with respect to the form of the data. There is no need of having an internal pattern in the response (conductivity). Nevertheless, if it exists, Random Forests will benefit of it.
- It is well-behaved with missing values. As we mentioned in the previous chapter, we might have some missing values on the input variables due to miss-measurements or other glitches. This is not an issue with this technique.
- Possibility of exploring the relations between variables. This may help us find some explanations and, in future works, solve the water treatment plant problem.
- Previous work in the field of regression of temporal series with Random Forests has obtained good results. See for example [9].

For completeness, we also explain in the following sections the tools needed to understand them better. We will start with the base of Random Forests, Decision Trees, a tree-based method. In section 3.2 we will explain a useful approach for reducing the intrinsic variance of Decision Trees, bagging or bootstrap aggregation. We end the chapter talking about Random Forests, from its principles and statistical properties, up to some key concepts such as variables importance measures.

3.1 Tree-Based Methods

The main idea of the tree-based methods for fitting is to use a "divide-and-rule" strategy. Working in the space of variables, or feature space, you divide into rectangles and fit each one with a simple model (constant, for example). When we have a new input, we check into what rectangle

it is and then the output would be the corresponding response. Regression trees can be traced back at least to 1963 with [14]. Nevertheless, we have followed the more modern approaches of Breiman [15] and Quinlan [16]. We are going to begin with an example in order to introduce the concepts in a smooth way.

3.1.1 A descriptive example

Let us consider a regression problem whose inputs are X_1 and X_2 and the output is $\mathbf{Y}$. We consider an Euclidean rectangle with the abscissae taking values of X_1 and the coordinates taking values of X_2. We choose some variable with certain split-point, optimising some value. Then one of both of these regions are split into two more regions, and this process is continued, until some stopping rule is applied. We end with an "if/else"-like structure where in each terminal statement (terminal node, or leaf) we fit the output. In each division we form *nodes* of the tree. Using a similar example that the one found in [17], we first split at $X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_3$ is split at $X_2 = t_4$. The result of this process is a partition into the five regions $R_1, \ldots, R_5$ shown in the right of figure 3.1. The corresponding

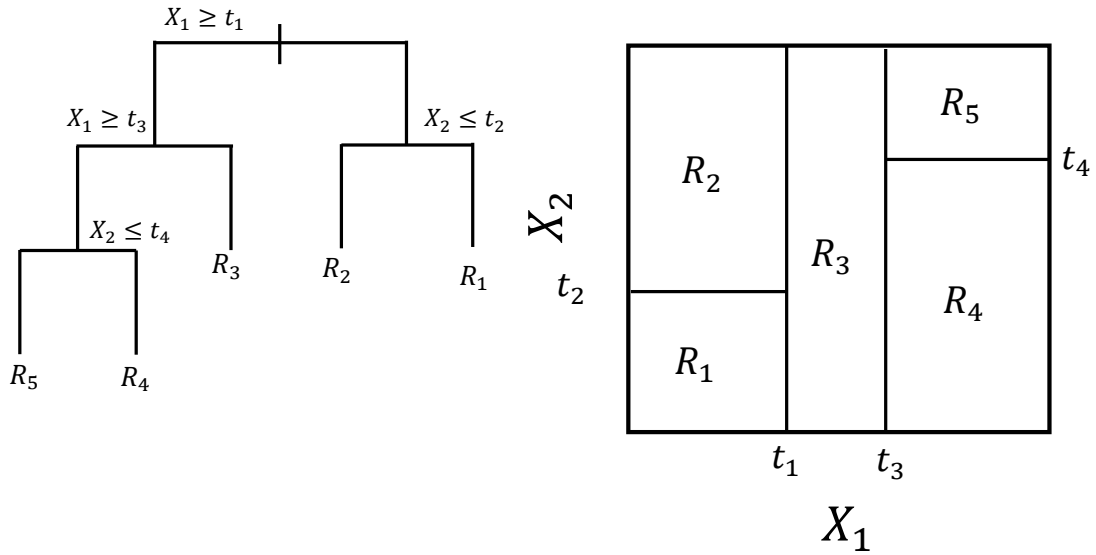

Figure 3.1: On the left, we can see the tree corresponding to the partition on the right. This partition on the right shows a two-dimensional feature space split recursively in a binary way applied to some fake data.

regression model predicts $\mathbf{Y}$ with a constant c_m in region R_m, formally

$$\hat{\mathbf{Y}}(X_1, X_2) = \sum_{m=1}^{5} c_m \mathbb{1}_{(X_1, X_2)}(R_m),$$

where $\mathbb{1}_x(A)$ is one if $x \in A$ and zero otherwise. A key advantage of the recursive binary tree is its interpretability. The feature space partition is fully described by a single tree, shown on the left of figure 3.1.

Let us now formalise the concept of *regression trees*. Let (x_i, y_i) for $i = 1, \dots, N$ be N pairs of observations where the input $x_i = (x_{i1}, \dots, x_{ip})$. We need to determine in an algorithmic way which variables to split and at which points. Suppose first that we have a partition into M rectangles $R_1, \dots, R_M$ and, as we have done in the example before we model the response as:

$$\hat{y} = \sum_{m=1}^{M} c_m \mathbb{1}_x(R_m),$$

for some c_m where the indicator function

$$\mathbb{1}_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A \end{cases}.$$

One reasonable strategy would be to minimise the sum of squares between the predictions and the real responses y_i. Then, it is trivial to see that the best $\hat{c}_m$ is the average of y_i in R_m:

$$\hat{c}_m = \overline{(y_i | x_i \in R_m)}. \tag{3.1}$$

But, computing this in every step is very expensive computationally, so we proceed with a greedy algorithm. Starting with the whole dataset, consider a splitting variable with index j and a split point s. We define

$$R_1(j, s) = \{X | X_j \leq s\} \qquad \text{and} \qquad R_2(j, s) = \{X | X_j > s\}.$$

The variable and splitting point are then found out by solving

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,2)} (y_i - c_2)^2 \right].$$

The c_i took their value in the inner minimisation following equation (3.1). For each variable, the split point s can be found in a very fast way, and hence also the pair (j, s).
Having found the best split, we partition the data into the two resulting regions and repeat the process in all the new split regions. This process makes the tree to grow in depth. We do it until some stopping criterion is applied, like the size of the rectangles.

But, rather than splitting the feature space in two at each step, we might consider to split it in more subspaces. While in certain occasions this could be a useful, it is not a good general, strategy. The problem with greater-than-binary splits is that the data is normally fragmented too much, leaving insufficient data at the next level down. Since this kind of splitting can be achieved by a series of binary splits, the latter are preferred.

This kind of trees are known as *classification and regression trees* or CART, and were formulated in [15]. As its name points out, this philosophy works out both for regression but also for classification. In that case, the objective would be assign each observation to a known group, or category. Although some of the algorithms are different, the way of establishing the methodology for solving the problem is very similar as we do with regression problems. In this project we will not enter in more details about classification problems since it is out of the scope of this work. General texts on classification include [18–21]. In [22] the reader can find a comparison of a large number of popular classifiers on benchmark datasets.
But these are not the only type of trees that exist. In order to correct some factors such as bias, conditional inference trees are introduced.

3.1.2 Conditional inference trees

One of the main problems of the CART approach is the bias towards variables with many possible splits. This issue was already diagnosed in [15, 23, 24] and appears due to the optimising splitting criterion over all possible splits previously mentioned. This bias appears also if many missing values in some variables exist, as is argued in [25].
In order to improve this flaw, the concept of *conditional inference trees* was introduced in [26], although inspired in works such as [27]. This concept was developed embedded in a unified framework using the theory of permutation tests developed in [28]. Some background on permutation tests implementation in recursive partitioning algorithms can be found in [29], but they focused only on special situations. Let us look a bit more deeper into the methodology of [26].

We aim to design an algorithm to fit a regression model, but describing at the same time the conditional distribution of an output given some input. This algorithm should be based on some learning sample $\mathcal{L}_n$, *i.e.*, a random sample of n independent and identically distributed (i.i.d.) observations. The approach of [26] is with some non-negative integer-valued weights $\mathbf{w} = (w_1, \ldots, w_n)$. If some observations are elements of the node of a tree, this would be represented with the corresponding weights being non-zero. Denoting the output by $\mathbf{Y}$ and by $\mathbf{X} = (X_1, \ldots, X_m)$ the vector of variables, the pseudo-code can be represented as:

1. Test the global null hypothesis H_0 of independence between any of the m variables and the response. If the hypothesis is rejected, select the j^*-th variable X_{j^*} with strongest association with $\mathbf{Y}$. For each variable X_j we will denote the partial null hypothesis by H_0^j.
2. Choose a subset A^* in order to determine two sub-vectors $\mathbf{w}_l$ and $\mathbf{w}_r$ where, for $i = 1, \ldots, n$

   $$w_{l,i} = w_i \mathbb{1}_{A^*}(X_{j^*,i})$$

   and $w_{r,i} = 0$ if and only if $w_{l,i} \neq 0$.
3. Repeat steps 1 and 2 modifying the weights until some stopping criteria is applied.

The first step is essentially an independence problem. A thorough and rigorous discussion about the topic can be found in [26], we outline here the conclusion. In this step we have to select X_{j^*} with

$$j^* = \underset{j=1,\ldots,m}{\operatorname{argmin}} p_j,$$

where p_j is the p-value of the conditional test for H_0^j. In other words, the variable with minimum p-value. The scaling from partial to global null hypothesis can be done in several ways. For simple approaches we refer to Bonferroni-adjusted p-values (see [30,31]); and for more advanced methods the multiple testing done in [32,33]. For an overview of computational methods for the conditional distribution and the p-value of the statistic that relates the output with the inputs, see [34]. Now, we reject H_0 when the minimum of the p-values is less than certain level α. This α may be seen therefore as a parameter for determining the size of the tree. Nevertheless, α can also be interpreted as a pre-specified level of the association test, and hence it controls the probability of wrongly rejecting H_0 per node. With this, to assure that every dependence is detected, we can increase this significance level. To avoid overfitting, a pruning of the tree could be done, *i.e.*, to remove the terminal nodes until the terminal splits are significant at certain level

$\alpha' \ll \alpha$. And, although we can choose α in a data-dependent way, with the classical convention of $\alpha = 0.05$ a good performance is demonstrated in [26].

Presented this way, conditional inference trees select variables in an unbiased way and the partitions induced by this recursive partitioning algorithm (introduced by a statistical approach) are not affected by overfitting.

Independently of the model tree used (CART or conditional inference trees), one of the advantages (perhaps the main one) of decision trees is their ability to deal with missing predictor values. Suppose our data has some missing predictor values in some or all of the variables (but *not* in the response variable). The first strategy might be to discard any observation with some missing values, but this could reduce an important amount the variable set (or *training set*). Another approach might be to impute (fill-in) the missing values, with some estimation such as the average of that predictor over the non-missing observations. But we need to know how this missing values replacement mechanism affects the observed data.

In plain language, data is *missing at random* if the mechanism resulting in its omission is independent of its unobserved value. Let us formalise this concept following the reasoning of [35]. Suppose $\mathbf{Y}$ is the response vector and $\mathbf{X}$ is the $N \times p$ matrix of inputs, some of which are missing. Denote by $\mathbf{X}_{\text{obs}}$ the observed entries in $\mathbf{X}$ and let $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, $\mathbf{Z}_{\text{obs}} = (\mathbf{X}_{\text{obs}}, \mathbf{Y})$. Let $\mathbf{R}$ be an indicator matrix with 1 at the i-th row and j-th column if X_{ij} is missing and zero otherwise. The data is said to be *missing at random* (MAR) if the distribution of $\mathbf{R}$ depends on the data $\mathbf{Z}$ *only* through $\mathbf{Z}_{\text{obs}}$, *i.e.*:

$$\Pr(\mathbf{R}|\mathbf{Z}, \theta) = \Pr(\mathbf{R}|\mathbf{Z}_{\text{obs}}, \theta),$$

where θ is some set of parameters. Data is said to be *missing completely at random* (MCAR) if the distribution of $\mathbf{R}$ does not depend neither on the observed nor on the missing data, *i.e.*:

$$\Pr(\mathbf{R}|\mathbf{Z}, \theta) = \Pr(\mathbf{R}|\theta).$$

Often the determination of which features are MCAR must be made from information about the data collection process.

There exist more types of missing value and imputation methods that try to correct their effects. For a more comprehensive review about missing data and imputation methods we refer the reader to [35–39].

But, for tree-based models, there are two better approaches for dealing with missing values. The first approach is making a new category for "missing". From this we might discover that observations with missing values for some measurement behave differently than those with non-missing values. This technique applies for categorical variables. The second more general approach is the construction of *surrogate* variables. When considering a predictor for a split, we use only the observations for which that predictor is not missing. Having chosen the best predictor and split point, we form a list of surrogate predictors and split points. The first surrogate is the predictor and corresponding split point that best mimics the split of the training data achieved by the primary split. The second surrogate is the predictor and corresponding split point that does second best, and so on. When we want to predict the output of some new observation, we use these surrogate splits in order. Surrogate splits exploit correlations between predictors to try and alleviate the effect of missing data. As one might imagine, the higher the correlation between the missing predictor and the other predictors, the smaller the loss of information due to the missing value.

Besides the fact that trees have a good handling with missing values in the predictor, as it happens in our applications, they are well-behaved in other aspects, such as:

- Natural handling of data of "mixed" type. We have not dealt with categorical data, since it is more natural in clustering or classification problems, but trees are also useful for these problems. Sometimes it can be useful to use categorical with continuous predictors: trees are effective also in these situations.
- Robustness to outliers in input space.
- Invariant under (strictly monotone) transformations of the individual predictors.
- Computational scalability. This is not an issue due to the aforementioned invariance.
- Ability to deal with irrelevant inputs. One of the reasons why this technique has become more popular is because no matter how many predictors you use in the tree at the beginning. If some of them are not important, the tree itself has no problem in dismissing them internally.

If the tree is small, it is quite interpretable the result. But, when the complexity arises, this interpretability might become compromised.

Nevertheless, one major problem with trees is their high variance and as a consequence its predictive power. Often a small change in the data can result in a very different series of splits, making the whole system ill conditioned. The major reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it. In order to try to correct this instability, in the next section we introduce the concept of bagging. With this particular form of aggregating trees, we can reduce the variance of the whole model.

3.2 Bootstrap aggregation

The methodology of Random Forest (further down explained) is based on grouping somehow several decision trees. In order to understand this, we first need to explain the concept of *bootstrap*, appeared for the first time in [40].

Suppose we have a model to fit (learn) a set of training data $\mathbf{Z} = (z_1, \ldots, z_N)$ where $z_i = (x_i, y_i)$. The basic idea of the bootstrap method is to randomly draw (extract) datasets with replacement from the training data, where each sample has the same size as the original training set. This is done B times. We will call these extractions *bootstrap datasets*. Then we use our model to refit each bootstrap dataset, and analyse the behaviour over the B replications.

Moreover, if we have some quantity $S(\mathbf{Z})$, from the bootstrap sampling we can estimate any aspect of its distribution. For example, its variance

$$\widehat{\mathbb{V}}\left[S(\mathbf{Z})\right] = \frac{1}{B-1}\sum_{b=1}^{B}\left(S\left(\mathbf{Z}^{*b}\right) - \overline{S}^{*}\right)^2,$$

where

$$\overline{S}^* = \sum_{b=1}^{B} \frac{S\left(\mathbf{Z}^{*b}\right)}{B}$$

being $\mathbf{Z}^{*b}$ the b-th bootstrap dataset. We can think of this variance as a Monte Carlo estimate of $S(\mathbf{Z})$ under sampling from the empirical distribution function for the data $\mathbf{Z}$. For a comprehensive revision of the bootstrap method, see [41] and [42].
Using this concept we can manipulate the bootstrap datasets for reducing variance and improving the accuracy of a fitting.

Let us consider a regression problem. Suppose we fit a model to our training data $\mathbf{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, and let $\hat{f}(x)$ be the prediction at input x. Bootstrap aggregation or *bagging* (see [43]) averages this prediction over certain subset of bootstrap samples. The immediate consequence of this is the reduction of the model variance.
For each bootstrap sample $\mathbf{Z}^{*b}$ for $b = 1, \ldots, B$, we fit the model, whose output is a prediction $\hat{f}^{*b}(x)$. The bagging estimate is defined by

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x). \tag{3.2}$$

Let us denote by $\widehat{\mathcal{P}}$ the empirical distribution, setting equal probability $(1/N)$ on each data point (x_i, y_i). The "real" bagging estimate is defined in terms of the expectation, instead of as an average. *I.e.*, by $\mathbb{E}_{\widehat{\mathcal{P}}}\left[\hat{f}^*(x)\right]$ where $\mathbf{Z}^* = \{(x_1^*, y_1^*), \ldots, (x_N^*, y_N^*)\}$ and each $(x_i^*, y_i^*) \sim \widehat{\mathcal{P}}$. With this set up, expression (3.2) is a Monte Carlo estimate of the true bagging estimate, approaching it as $B \to \infty$.

In our framework, $f(x)$ denotes the decision tree's prediction at input vector x. Each bootstrap tree will typically involve different features (input variables) than the original, and might have a different number of terminal nodes. The bagged estimate is the average prediction at x from these B trees. As we said, by doing this, bagging can dramatically reduce the variance of unstable procedure like trees, leading to improved prediction. Let us proof this previous affirmation:

Assume our training set $\{(x_i, y_i) \colon i = 1, \ldots, N\}$ is formed by independent samples from a distribution $\mathcal{P}$, and consider the ideal aggregate estimator $f_{\text{ag}}(x) = \mathbb{E}_{\mathcal{P}}\left[\hat{f}^*(x)\right]$. The bootstrap dataset is $\mathbf{Z}^* = \{(x_i^*, y_i^*) \colon i = 1, \ldots, N\}$ sampled from $\mathcal{P}$. Since this estimation is from the actual population rather than a sample, we can not use it in practice, but it is useful for analysis. Thus, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}}\left[Y - \hat{f}^*(x)\right]^2 &= \mathbb{E}_{\mathcal{P}}\left[Y - f_{\text{ag}}(x) + f_{\text{ag}}(x) - \hat{f}^*(x)\right]^2 \\
&= \mathbb{E}_{\mathcal{P}}\left[Y - f_{\text{ag}}(x)\right]^2 + \mathbb{E}_{\mathcal{P}}\left[\hat{f}^*(x) - f_{\text{ag}}(x)\right]^2 \\
&\geq \mathbb{E}_{\mathcal{P}}\left[Y - f_{\text{ag}}(x)\right]^2,
\end{aligned}$$

due to the aforementioned independence. The last inequality comes from the variance of the estimator $\hat{f}^*(x)$ around its mean $f_{\text{ag}}(x)$. Hence, true population aggregation never increases this mean squared error. This suggests that bagging will often decrease mean squared error.

3.3 Random forests

With all the previous knowledge we can now explain the main methodology followed in this project. *Random forests* is a substantial modification of bagging that builds a large collection of *de-correlated* trees, and then averages them. They are a powerful tool, and simple to train and tune the model parameters.
The early development of Breiman's [44] notion of random forests was influenced by the work of Amit and Geman [45] who introduced the idea of searching over a random subset of the available decisions when splitting a node, in the context of growing a single tree. The idea of random subspace selection from Ho in [46] was also influential in the design of random forests. This same author was who proposed the term of "Random Forest" in [47]. In this method a forest of trees is grown, and variation among the trees is introduced by projecting the training data into a randomly chosen subspace before fitting each tree or each node. Finally, the idea of randomised node optimisation, where the decision at each node is selected by a randomised procedure, rather than a deterministic optimisation was first introduced by Dietterich (see [48]).

In [44] we can find a description of a method of building a forest of uncorrelated trees, combined with randomised node optimisation and bagging. In addition, this paper combines several ingredients, some previously known and some novel, which form the basis of the modern practice of random forests, in particular:

- Using out-of-bag error as an estimate of the generalisation error.
- Measuring variable importance through permutation.

As the reader must have already observed, the essential idea in bagging is to average many noisy models, and hence reduce the variance. Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Trees are notoriously noisy, and therefore they benefit greatly from the averaging. Moreover, since each tree generated in bagging is identically distributed (i.d.), the expectation of an average of B such trees is the same as the expectation of any one of them. This means that the bias of bagged trees is the same as that of the individual trees, and the only hope of improvement is through variance reduction.

An average of B i.i.d. random variables, each with variance σ^2, has variance $\frac{1}{B}\sigma^2$. If the variables are simply i.d. with positive pairwise correlation ρ, the variance of the average (see [17]) is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{3.3}$$

As B increases, the second term vanishes, but the first remains, and hence the size of the correlation of pairs of bagged trees limits the benefits of averaging. The idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables. Before each split, select $m \leq p$ of the input variables at random as candidates for splitting. Typically values for m are $\sqrt{p}$ or even as low as 1, according to [17]. Nevertheless, the inventors make the recommendation that, for regression, the default value for m is $\lfloor p/3 \rfloor$ and the minimum node size is five. After B such trees $\{T(x;\Theta_b)\}_{b=1}^{B}$ are

grown, the random forest prediction for regression is

$$\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b). \tag{3.4}$$

Here, Θ_b characterises completely the b-th random forest tree (split variables and points, terminal nodes...). The pseudocode of all this process can be summarised as follows:

for $b = 1$ **to** B: **do**
  Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
  Grow a tree T_b to the bootstrapped data:
  while Node size $> n_{\min}$ **do**
    Select m variables at random from the p variables.
    Pick the best variable/split-point among the m.
    Split the node into two daughter nodes.
  end while
end for
Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.
To make a prediction at a new point x:

$$\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

An important feature of random forests is their use of *out-of-bag* (OOB) samples. For each observation $z_i = (x_i, y_i)$, the idea is to construct its random forest predictor by averaging *only* those trees corresponding to bootstrap samples in which z_i did not appear. These are the so called OOB samples. It can be shown (see [17]) that once the OOB error stabilises, the training (repeating the fitting) can be terminated.

Up to this point we have not dealt with one of the main problems of fitting methods: *overfitting*, when the model predicts properly for the training set at the expense of a poor quality prediction for new datasets. When the number of variables is large, with few relevant variables, random forests are likely to perform poorly with small m. This is because at each split, the chance of the relevant variables to be selected could be small. In opposition, when the number of relevant variables increases, the performance of random forests is robust even if more noise variables are added, surprisingly.

Regarding the overfitting issue, it is certainly true that increasing B does not cause the random forest sequence to overfit; like bagging, the random forest estimate (3.4) approximates the expectation

$$\hat{f}_{\mathrm{rf}}(x) = \mathbb{E}_{\Theta}\left[T(x;\Theta)\right] = \lim_{B\to\infty} \hat{f}(x)_{\mathrm{rf}}^{B}. \tag{3.5}$$

The distribution of Θ is conditional on the training data. However, this limit can overfit the data, since the average of fully grown trees can result in a too rich model, and incur unnecessary variance. In [49] small gains in performance are demonstrated by controlling the depths of the individual trees in random forest.
The explanation of the Random Forest's resistance to overtraining can also be found in Kleinberg's theory of stochastic discrimination (see [50, 51]).

Some last remarks about the formalisation of the de-correlation effect and the bias must be done.
Using equations (3.3) and (3.5) we see that the ideal variance is

$$\mathbb{V}\left[\hat{f}_{\mathrm{rf}}(x)\right] = \rho(x)\sigma^2(x),$$

where $\rho(x)$ is the correlation induced by the sampling distribution of $\mathbf{Z}$ and θ between any pair of trees used in the averaging:

$$\rho(x) = \mathrm{corr}\left[T\left(x;\Theta_1(\mathbf{Z})\right), T\left(x;\Theta_2(\mathbf{Z})\right)\right].$$

Here $\Theta_1(\mathbf{Z})$ and $\Theta_2(\mathbf{Z})$ are a randomly drawn pair of trees extracted from $\mathbf{Z}$; $\sigma^2(x)$ is the sampling variance of any single randomly drawn tree:

$$\sigma^2(x) = \mathbb{V}\left[R\left(x;\Theta(\mathbf{Z})\right)\right].$$

This variability is both a result of the sampling variability of $\mathbf{Z}$ itself; and conditional on $\mathbf{Z}$, because of the bootstrap and feature sampling at each split.
Moreover, the conditional covariance of a pair of tree fits at x vanishes, since the bootstrap and feature sampling are i.i.d..

As for the bias (as in bagging), the bias of a random forest is the same that the bias of any of the individual sampled trees $T(x;\Theta(\mathbf{Z}))$. Formally:

$$\begin{aligned}\mathrm{Bias}(x) &= \mu(x) - \mathbb{E}_{\mathbf{Z}}\left[\hat{f}_{\mathrm{rf}}(x)\right] \\ &= \mu(x) - \mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}_{\Theta|\mathbf{Z}}\left[T\left(x;\theta(\mathbf{Z})\right)\right]\right],\end{aligned}$$

where we have made explicit the dependence of Θ with respect to $\mathbf{Z}$. Because of the restrictions imposed by the randomisation and the reduced amount of samples, this is usually greater than the bias of a tree grown from $\mathbf{Z}$. Therefore, the improvements in prediction of random forests are uniquely as a result of variance reduction. Due to the equations derived before, any discussion of bias depends on the unknown tree function. But in practical terms (although for different models the bias may differ) usually as m decreases, the bias increases.

Before moving on to the next chapter, where we will present the results, we have considered explaining the concept and the different ways of measuring variable importance. Since it will be a key concept in the next chapter, we end this one dedicating a subsection to it.

3.3.1 Variable importance measures

Now we will try to quantify the importance of each input variable of a random forest. The broad idea is that a variable is important if when it is removed from the algorithm, the accuracy of the prediction diminishes. Reciprocally, a predictor will be irrelevant if the accuracy of the prediction does not change if it is dismissed. Predictors are not usually equally relevant. With a variable important measure (VIM from now on) we can rank variables and identify those which are more influential to the prediction. There is not a unique measure for variable importance: moreover, in some of the cases the existent ones are not even equivalent [52].

We will now outline the most relevant and widely used VIMs for random forests. Namely, Gini VIM, permutation VIM (based on CART and conditional inference trees), conditional permutation VIM, variable selection based on tree-based concept of minimal depth statistic and the recent intervention in prediction measure. All the measures we will explain will be for regression. Although in some cases are similar, for classification other VIMs apply. For a comprehensive review of these measures (with the exception of the intervention in prediction measure) see [53].

Gini VIM, or GVIM. Proposed in [44], GVIM is based on the concept of impurity measure of a tree node. The node impurity is defined as the residual sum of squares (for regression). In every split this quantity will be smaller (this is related with the overfitting issue). Hence, the importance of a variable is defined as the total decreasing of these node impurities at every split of the variable chosen, averaged over all the trees. As pointed out and reasoned in [54], when there are different types of variables, GVIM is biased towards continuum variables or with many categories (in the case of categorical variables).

Permutation VIM, or PVIM. Formalised and studied in [55], the rationale under the importance of a variable is the following. By randomly permuting the predictor variable, its original association with the response is broken. When the permuted variable is used to predict the response, the prediction accuracy decreases substantially, if the original variable was associated with the response. Thus, a reasonable measure for variable importance is the difference in prediction accuracy before and after the permutation. With this VIM, the bias existent in GVIM is corrected, as proved in [55].

Conditional permutation VIM, or CPVIM. One issue with PVIM is that a bias appears when there exist correlated variables. This was corrected in [56] where they proposed a conditional permutation scheme. This measure is computed conditionally on the values of other associated predictor variables (possible confounders). In order to get a clearer understanding of this concept, let us use an example from [57]. Although this example is with categorical variables, the idea behind it is the same. Let $\mathbf{Y}$ be a binary response where $Y = 0$ means that a diagnosis about fetal health during pregnancy is incorrect and $Y = 1$ means the opposite. Let us consider the following predictors: X_1, which asseses the quality of ultrasound devices at the hospital; X_2 which reflects if the hospital staff are trained to use them and interpret the images; and X_3, which establishes the cleanliness of hospital floors. The second and third variables are correlated since they are related to the quality of the hospital, but the important one is X_2. So, conditionally on X_2, X_3 does not have any effect on $\mathbf{Y}$. This distinction is achieved in CPVIM, but not in PVIM. According to [58], CPVIM is more appropriate if the aim is to identify the influential variables without considering the correlated effects.

Variable selection based on tree-based concept of minimal depth statistic, or varSelMD. The minimal depth statistic (MD) establishes the capacity of prediction of a variable by its distance to the root node of a tree, where the first split occurs. A smaller value corresponds to a more predictive variable (see [59]). More into details, we define a *maximal subtree* for a predictor X_j as the largest subtree whose root node is split using X_j. In other words, no other parent node of the subtree is split using X_j. The MD of a variable X_j is the shortest distance from the root of the tree to the root of the closes maximal subtree of X_j. A method based on this concept of MD for variable selection was introduced in [60]. It uses all variables at once. Variables with an average MD exceeding the average MD threshold are considered noisy and hence dismissed from the final model.

Intervention in prediction measure, or IPM. Proposed in [61], this measure can be

computed whether with new data or with data used in the training set. Depending on the case, the methodology is different. For computing the IPM, the new case is added to all the trees in the forest. For each tree, the cases goes through a series of nodes. The variable split in these nodes is recorded. The percentage of times a variable is selected along the case's way from root to the terminal node is calculated for each tree. Then, the IPM is obtained by averaging those percentages over the whole set of trees. Notice that the IPM of a new case does not need to know the true response with this data. In the case of using data present in the training set, the IPM is just computed averaging only over the trees where the case belongs to the OOB set. Due to its formulation, IPM can be computed for subsets of data, without needing of regrowing the forest for those subsets.

A summary of this measures can be found in table 3.1, extracted from [62].

Measure	Main references	Key characteristic	Tree based on
GVIM	[63]	Node impurity	CART
PVIM	[26, 44, 54]	Accuracy after variable permutation	CART and CIT
CPVIM	[56]	Alternative of PVIM: Conditional permutation	CIT
varSelMD	[59, 60]	Variable selection based on MD	CART
IPM	[61, 62]	Variables intervening in prediction	CART and CIT

Table 3.1: Summary of the VIMs previously described. The columns respectively mean: the VIM, the main work(s) about the topic in the literature, the broad idea behind it and if it applies to CART-based random forests or conditional inference trees (CIT)-based random forests. Table extracted partially from [62].

In the next chapter we will see the applications of some of these methods and VIMs for our problem, as well as some useful representations of these results.

Chapter 4

Results

In this chapter we will apply the previously learned techniques to the sea water infiltration problem. As we previously said, the goal is trying to predict FACSA's sewage conductivity measures with the data provided by them and the datasets obtained from [13].

We will start explaining the software we have been using, followed by some preliminary analysis of the data. We will use then Random Forests for extracting some conclusions, measuring the errors in the prediction. In section 4.4 we measure some variables importance, using several justified scenarios. We end the chapter presenting a useful visualisation tool for the prediction, proximity plots.

4.1 Software considerations

Before presenting the results, let us expose some software considerations. The software we have been used has been the `R` programming language [64]. We have chosen it because it is open source, and hence all the code is reproducible, and because it is oriented to statistical computing.

For completeness, let us enumerate the main packages that could have been used for our purposes. Breiman's Random Forest algorithm [44] is implemented in the `R` package `randomForest` [65] and also in the `R` package `randomForestSRC` [66]. Random Forests based on conditional inference trees can be found in the `R` package `party` [56]. An aspect we have not considered up to now is the concept of *multivariate random forests*, when several responses are allowed. These techniques have been gaining more popularity in the past few years [67]. Multivariate Random Forests can be computed by the `R` package `randomForestSRC` and the `R` package party, but not by the `R` package `randomForest`. In the `R` package `randomForestSRC`, for multivariate regression responses, a composite normalised mean-squared error splitting rule is used; for multivariate classification responses, a composite normalised Gini index splitting rule is used (see [68]); and when both regression and classification responses are detected, a multivariate normalised composite split rule of mean-squared error and Gini index splitting is invoked.

We prooced now with the variable importance measures explained in section 3.3.1. GVIM

and PVIM are derived from Random Forests based on CART and can be computed with the `R` package `randomForest`. In this case, PVIM is scaled (normalized by the standard deviation of the difference) by default in the `randomForest` function from the `R` package `randomForest`. The problems of this scaled measure are explored by [69]. PVIM can also be derived from based on conditional inference trees and obtained with the `R` package `party`. CPVIM is based on conditional inference trees, and can be calculated using the `R` package `party`. The procedure for variable selection based on the tree-based concept termed MD proposed by [60] (varSelMD) is available from the `R` package `randomForestSRC`.

Here we have mostly used random forests based on conditional inference trees, due to the bias correction explained in the previous chapter.
For the decision trees we have used the `ctree` instruction of the `R` package `party`. Random Forests have been computed with the `cforest` and `predict` functions of `party` package. With respect to the VIMs, we have used `varimp` from `party` package and `ipmparty` from the `R` package `IPMRF` [62]. The distance between variables (outlined below) has been plotted using the `proximity` instruction from `party` package and the function `cmdscale` from the `R` package `stats` [64]. Unless stated otherwise, all the optional parameters used are the default ones of the function concerned.

4.2 Preliminary analysis

As a first inspection we look graphically at the conductivity values. We have values from the first day of January, 2012, to the last day of May, 2017.

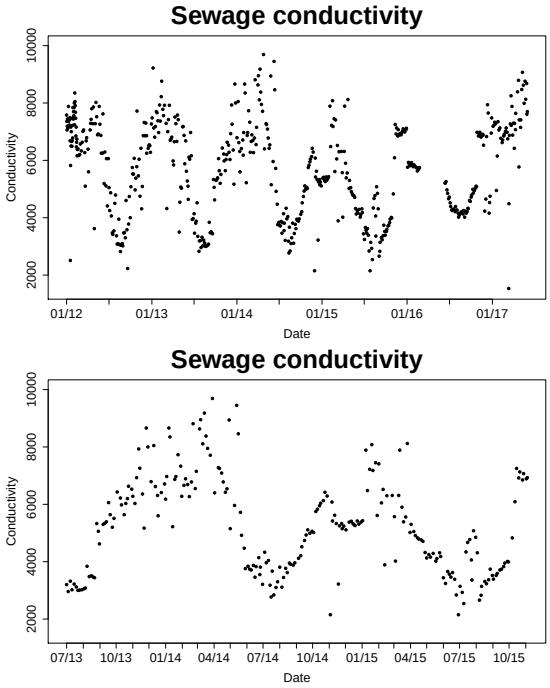

Figure 4.1: Plots of the water conductivity measured by FACSA. The upper plot corresponds to the whole dataset, while the figure below is a zoom of it. Dates are in format mm/yy.

In figure 4.1 we can see this dataset, zoomed-in below. We remark some features in these plots:

- The periodicity. This is the reason of the enhancing, in order to notice an inner pattern that seems to be annual. We can see how the conductivity rises in the winter months and decreases in the summer months.
- The missing values, especially since 2016. This is a problem, since Random Forests deal with missing values *only* in the non-response variables.
- The large amount of (presumed) outliers. This is not an issue since, as we discussed in the previous chapter, the technique used is well-behaved with outliers.

For a good visualisation of what is happening we decided to begin with a prediction with a decision tree. We take as predictors the water inflow and accumulated rain (since these were the first predictors given by FACSA) for the three previous days[1] to the day we want to predict. The number after the variable's name indicate how many days before the measure was taken. *I.e.*, Flow1, Flow2... correspond to the water inflow of "yesterday", the day before yesterday and so on.

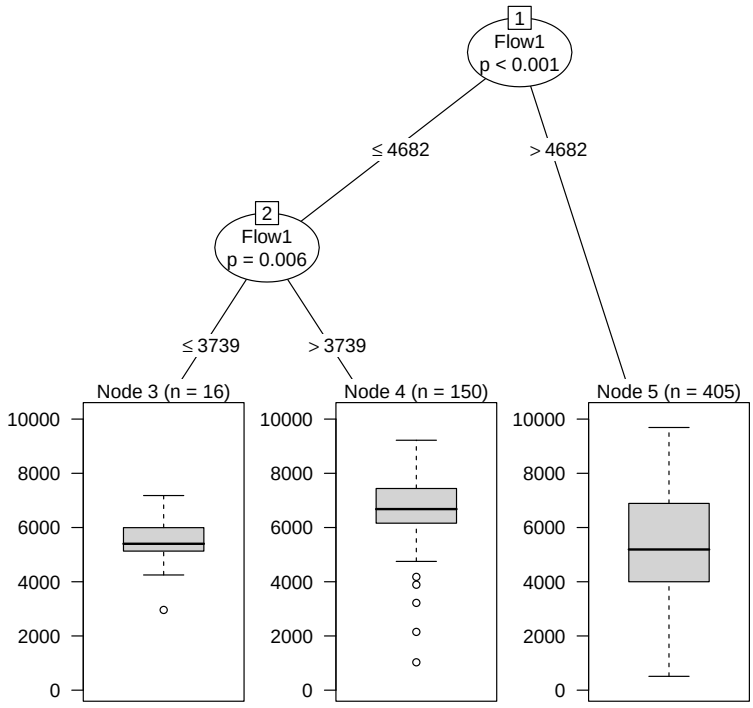

Figure 4.2: Regression tree for conductivity using water flow and accumulated rain for the three previous days.

[1] Actually we take the three previous data. But since we have a large amount of points, we are confident that this is not an issue.

In figure 4.2 we notice that not only the rains do not have influence, but neither the previous days to the first one. In the next section we will use random forests to add more variables, measure errors and further down check variable importance.

4.3 Predictions with Random Forests

For the case of using Random Forests we have set up several scenarios using different predictors:

- The ones considered previously, *i.e.*, water inflow and accumulated rain together with rain intensity. We have used these values for the three and ten previous days. We thought these might be influential in order to characterise the periods where torrential rain is present. As we said before, this intensity has been computed as the maximum rainfall divided by ten (since the measures were ten minutes long).
- Only the average temperature for the three and ten previous days. The temperature usually also behaves periodically with higher values in summer and lower in winter months. This may help predict the conductivity results.
- Only the conductivity itself for the three and ten previous days. Since in plots 4.1 we have observed a latent pattern, we could use this as a predictor as well. The rationale behind is to think that if the previous days the conductivity was growing, today it will probably grow too.
- The conductivity for the previous year's week. *I.e.*, if we want to predict the conductivity for 04/10/2017 we use as predictor the values of conductivity (if available) from 01/10/2016 to 07/10/2016. The idea is to mimic the behaviour of the previous year with some error margin (plus-minus three days).
- All the previous predictors using (if possible) three days and using ten days.

As for the error measures, we have considered two of the most popular measures that could provide some useful information about the prediction: residual sum of squares (RSS), computed as the sum of the squared residuals and the maximum residual in absolute value (Max. error). In table 4.1 we can see all these cases summarised, where "FH" is the abbreviation for "FACSA's hypotheses", meaning when using as predictors accumulated rain and water inflow.

Predictors		Error measures	
Variables	Time-window	RSS	Max. error
FH+Rain intensity	3 previous days	1.36×10^9	5.84×10^3
	10 previous days	1.29×10^9	6.25×10^3
Avg. temperature	3 previous days	8.09×10^8	5.76×10^3
	10 previous days	7.41×10^8	6.02×10^3
Conductivity	3 previous days	5.48×10^8	6.73×10^3
	10 previous days	5.16×10^8	7.16×10^3
	Last year's week	1.07×10^9	5.86×10^3
Everything	3 previous days	5.27×10^8	6.73×10^3
	10 previous days	5.47×10^8	6.94×10^3

Table 4.1: Some error measures of several scenarios explained before. Grey cells denote the minimum value for the corresponding column.

We notice in this table some surprising results, to say the least. As happened before, adding as variables rain values does not seem to improve the prediction, even if we use rain intensity as well. The temperature seems to be important, since we achieve the minimum maximum error when using the three previous days. As for the conductivity, we can see how it has been a good choice since we achieve the lowest RSS but the information from the previous year does not improve this prediction. When using the whole dataset of predictors we do not achieve better results.
Comparing only between the time-window used, using three days instead of ten always improves the Max. error and using ten always improves the RSS, with the exception when using everything. In the next section we will investigate the importance of the variables depending also on the previous days used. This could be useful in order to decide which threshold to establish when using previous days.

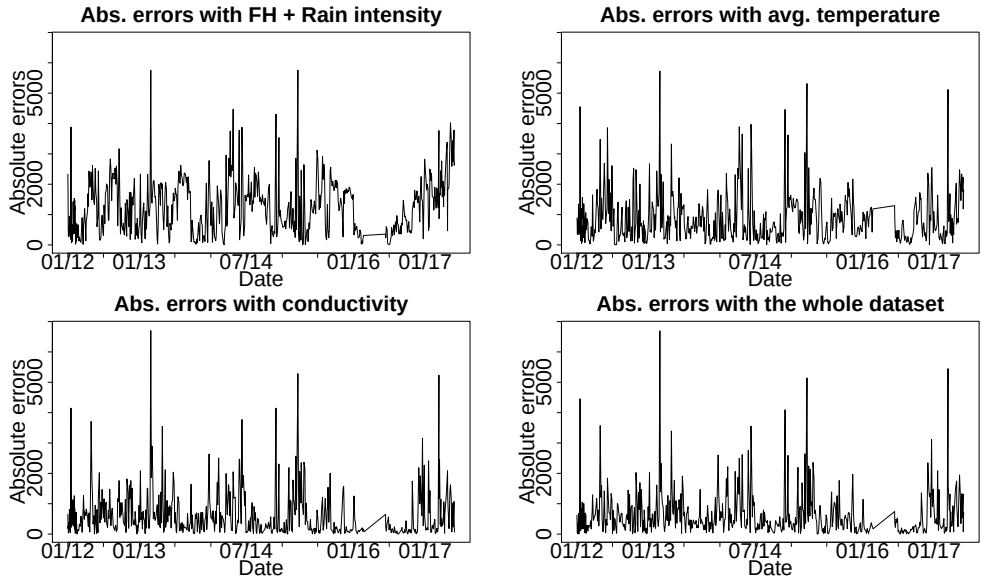

Figure 4.3: Different error plots depending on the predictors used for the predictions.

In figure 4.3 we can see the different error plots depending on the predictors used for the predictions. In all of the cases we have plotted them using the three previous days. We check how the peaks in the errors corresponds with the outliers in conductivity (see figure 4.1), more precisely with the conductivity drops. In general terms we see how, also here, the error with rain data and water inflow is higher than with conductivity.
As a final comment about these measures, the magnitudes should not surprise us since the conductivity's values vary between 2×10^3 and 1×10^4 approximately.

4.4 Variable importance measures

We now check for some of the VIMs presented in section 3.3.1. Since the fittings we have done with Random Forests are based on conditional trees, we will not use neither GVIM nor varSelMD (see table 3.1) because they are only available for CART-based random forests. Moreover, we will not use CPVIM neither because the `R` implementation uses a considerable amount of memory storage, and the computation time exceeded our expectations.

In light of the errors showed in the previous section (see table 4.1) and the FACSA's hypotheses we have fixed the predictors to water flow, average temperature, rain intensity, accumulated rain and distinguish between and adding and removing conductivity.

In the case of using PVIM with conductivity we obtain the plot shown in figure 4.4. Once

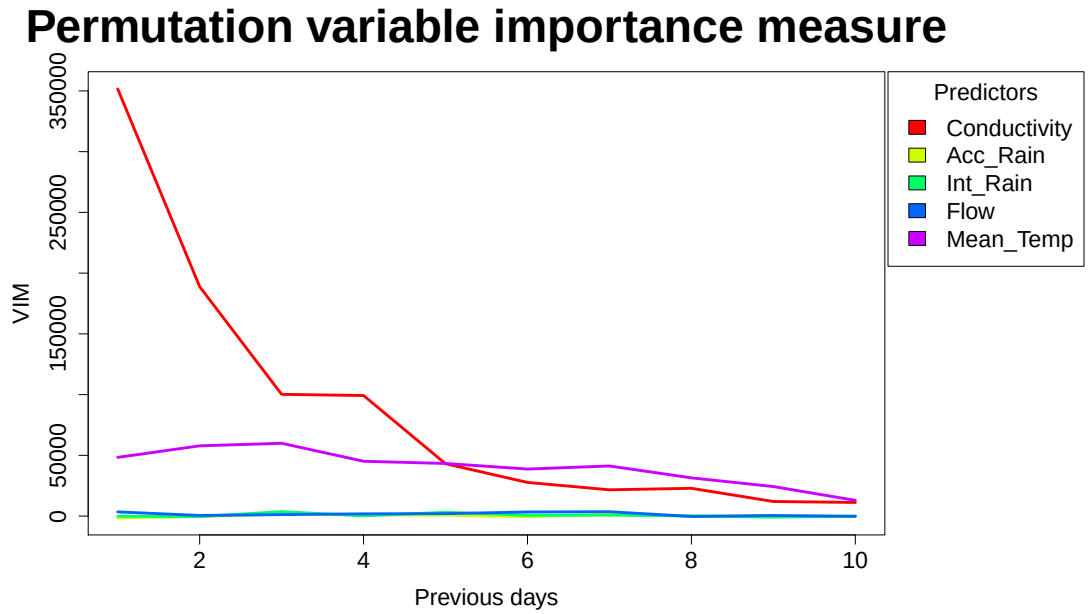

Figure 4.4: VIM depending on the number of previous days used.

again, the conductivity is the most relevant variable almost with independence of the day. In this same variable we observe a quasi-monotone decreasing, meaning that the less recent the day

used, the less important it is. Regarding the other predictors, we can not distinguish a clear tendency, maybe due to random fluctuations. Pluvial variables are barely significant.

We see how this results do not present contradiction with the VIM measured with IPM, as shown in figure 4.5.

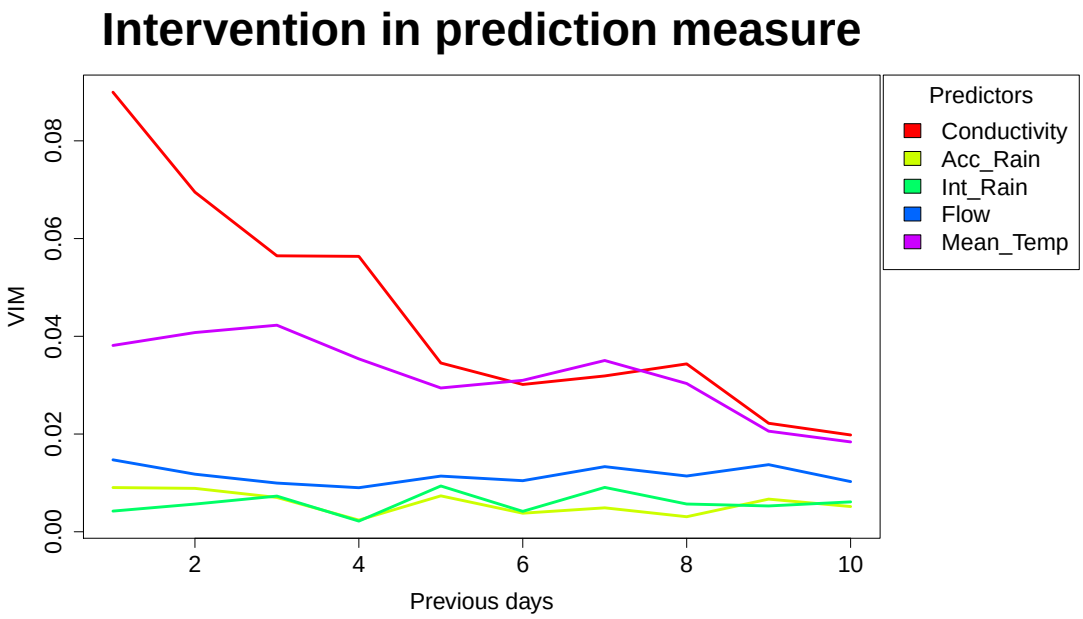

Figure 4.5: VIM depending on the number of previous days used.

The main difference in this case is that the distance between the importance of conductivity and mean temperature is shrunk. Here we can see too a decreasing tendency in the mean temperature, but water flow, accumulated rain and rain intensity are not relevant either.

As a new case we have decided to repeat this measures but removing the conductivity as predictors. The results can be seen in figures 4.6 and 4.7.

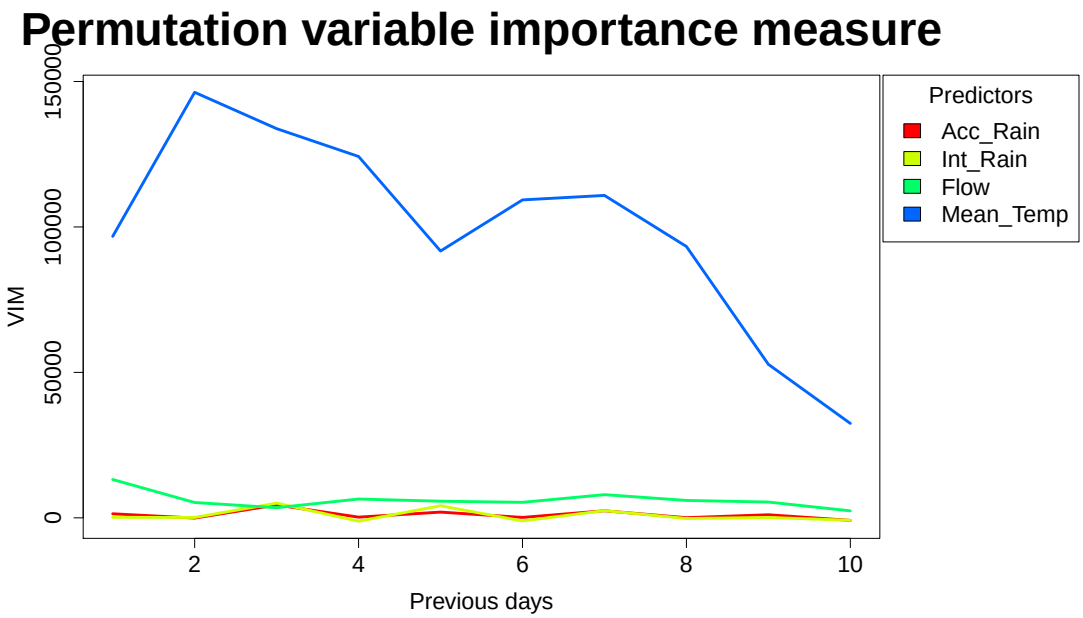

Figure 4.6: VIM depending on the number of previous days used.

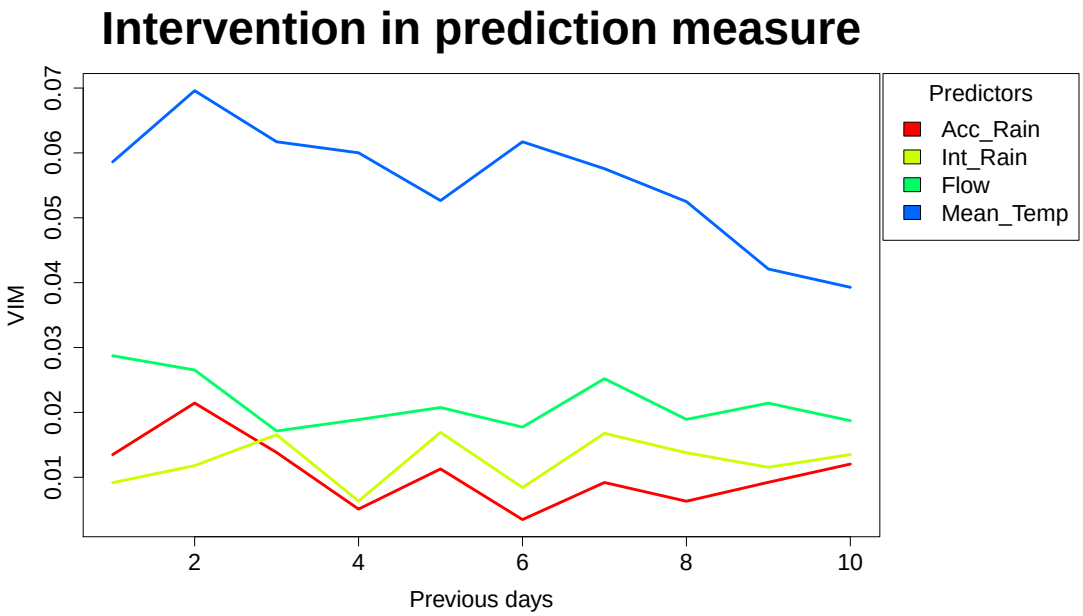

Figure 4.7: VIM depending on the number of previous days used.

As we could have expected, in this case the meaningfulness of the mean temperature is

enhanced, reflecting the aforementioned decreasing tendency. Nevertheless, the other predictors have not increased their importance.

4.5 Proximity plot

Besides variable importance measures a random forest offers a helping tool for data interpretation: a proximity plot. The *proximity* between two cases i and j is defined as the proportion of trees where cases i and j are in the same terminal node. These proximities can be assembled in a matrix in order to visualise it.

Let D be the *dissimilarity matrix* whose element d_{ij} for $i, j = 1, \ldots, N$ is built as one minus the proximity between cases i and j. Let M be the matrix with elements $m_{ij} = -0.5 \cdot d_{ij}^2$ and

$$B = (I_N - N^{-1}U)M(I_N - N^{-1}U) \tag{4.1}$$

where I_N is the $N \times N$ identity matrix and U is the $N \times N$ matrix with all the elements equal to one.
The matrix D is an Euclidean distance matrix (*i.e.*, a matrix representing the spacing of a set of N points in Euclidean space), only if B is positive semidefinite (see Theorem 14.2.1 in [70]). If this is the case, we can use *principal component analysis* (PCA) in order to represent them in at most $N - 1$ dimensions (Theorem 14.4.1 in [70]). With PCA we obtain a set of points such that distances between the points are approximately equal to dissimilarities. We could be in the case where the aforementioned matrix is not an Euclidean distance matrix. We can still apply PCA but only as an approximation (Theorem 14.4.2 in [70]).

We have decided to make this representation for the two best cases obtained, *i.e.* using only conductivity of the past ten days and using only average temperature of the past three days (see table 4.1).
In neither of the two cases the matrix B of equation (4.1) is semidefinite positive: the eigenvalues are approximately between $-8 \cdot 10^{-4}$ and $5 \cdot 10^{-5}$ being non-negative between the $41\% - 57\%$. So we must interpret the proximity plots as approximations. These diagrams are shown in figure 4.8. We have coloured and shaped the points according to the season: blue for winter, green for spring,

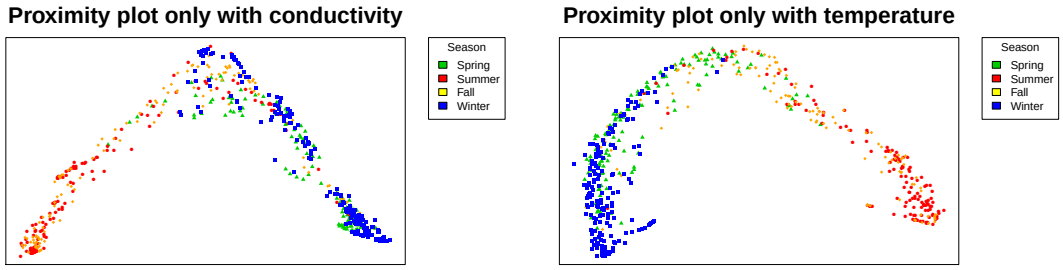

Figure 4.8: Proximity plots using the two best predictors according to table 4.1, distinguishing between seasons.

yellow for fall and red for summer. We can observe in both cases a horsehoe effect, which is

usually observed when there is an intern ordering of the data [71]. As it happened in [61], in our case it is because the seasons are ordered factors, ranging from winter to summer while fall and spring lie in between. We notice how there is not a clear clustering of these latter seasons, probably because temperatures are smoother due to the Mediterranean climate, and hence less distinguishable from winter and summer.

In the next chapter, we will end giving some last remarks about what has been accomplished in this master thesis and we will point out some possible lines of future work.

Chapter 5

Final remarks

The work described in this document has been a project funded by a research internship of the UJI's *Cátedra FACSA de Innovación del Ciclo Integral del Agua* [8]. Let us end this work with some finals considerations.

FACSA (see chapter 2) detected the problem of saltwater infiltration to their sewage treatment plant. This may cause serious damage to the water filters and hence to the quality of their service. As a marker of this leakage, they considered the measure of water electrical conductivity.

In order to analyse and predict these conductivity measures we have decided to use the technique of Random Forests (see chapter 3). As well as its flexibility with missing values, one key characteristic of Random Forests is its ability to measure the importance of the variables, the predictors. This was presented embedded in a statistical benchmark, explaining previously all the needed knowledge in the mentioned chapter 3.

Some interesting results have been dropped from our analysis. The main one is the absence of evidence indicating that the rain affects the change in conductivity. In light of this, we have decided to use other predictors such as the temperature, the water inflow or the conductivity in past days itself. The most relevant by far has resulted to be the conductivity, which makes sense due to the inner pattern shown in figure 4.1. If we do not take this into account, the temperature (average temperature) is the most important. Nevertheless, we remark that we have not talked about causation, since with the tools that we have at our disposal, only correlations are possible.

5.1 Future work

Some lines of future research have been opened along this report. We now summarise them:

- Tuning the parameters of the model. We have used the `R` software default values, and maybe a better fitting could be achieved choosing some other. Nevertheless, according to [44] the only adjustable parameter to which random forests are somewhat sensitive is the number of variables randomly sampled as candidates at each split.

- Adding some other predictors we have not considered. Tide's data might be good indicators, since maybe when the sea rises there is a higher probability of a sea water infiltration. We have not considered them here due to the inability to find some easily accessible and manipulable data.
- Finding the causes, beyond finding correlations. In several works, (see for example [72]) they consider climate change. A hypothesis might be that due to the climate change, the sea level rises and more infiltration occurs. We have found that the temperature is key, but we do not have enough evidence to assure any causation.
- Changing the main technique. Since this is a preliminary work, we decided to use one of the most flexible ones, but maybe other approaches are more adequate. Options might be to use temporal series or functional data analysis [73, 74].
- Using some other measures to quantify saltwater infiltration. There is some material in the literature such as a technical report [6] and a master's thesis [75] indicating that in some cases there is not a clear correlation between conductivity and sea water infiltration.

Bibliography

[1] "Ficha EDAR," Sep. 1993, in Spanish. Accessed: 17 Oct. 2017. [Online]. Available: http://www.epsar.gva.es/sanejament/instalaciones/edar.aspx?id=44

[2] "Los vertidos fecales en la ría del Sella se deben a infiltraciones de agua marina," Mar. 2016, in Spanish. Accessed: 07 Nov. 2017. [Online]. Available: http://www.lne.es/oriente/2016/03/15/vertidos-fecales-ria-sella-deben/1897182.html

[3] L. Simonsson, Å. G. Swartling, K. André, O. Wallgren, and R. J. Klein, "Perceptions of risk and limits to climate change adaptation: Case studies of two swedish urban regions," in *Climate change adaptation in developed nations*. Springer, 2011, pp. 321–334.

[4] M. Schirmer, S. Leschik, and A. Musolff, "Current research in urban hydrogeology–a review," *Advances in Water Resources*, vol. 51, pp. 280–291, 2013.

[5] J. F. Flood and L. B. Cahoon, "Risks to coastal wastewater collection systems from sea-level rise and climate change," *Journal of Coastal Research*, vol. 27, no. 4, pp. 652–660, 2011.

[6] J. Phillips, "Saltwater intrusion and infiltration into king county wastewater system," 2011.

[7] "Plan especial de alerta y eventual sequía en la confederación hidrográfica del júcar," Mar. 2007, in Spanish. Accessed: 7 Nov. 2017. [Online]. Available: http://www.chj.es/es-es/medioambiente/gestionsequia/Documents/Plan%20Especial%20Alerta%20y%20Eventual%20Sequia/PES_Marzo_2007.pdf

[8] "Grant call for research intern," Feb. 2016, in Spanish. Accessed: 10 Oct. 2017. [Online]. Available: http://www.uji.es/seu/info-adm/tao/convocatoriesRH?pArId=24650&pCategoria=4&pSubCategoria=9

[9] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz, "Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks," *BMC bioinformatics*, vol. 15, no. 1, p. 276, 2014.

[10] "Facsa Ciclo Integral del Agua," Apr. 1997, in Spanish. Accessed: 10 Oct. 2017. [Online]. Available: http://www.facsa.com

[11] "Grupo Gimeno," Mar. 2001, in Spanish. Accessed: 10 Oct. 2017. [Online]. Available: http://www.grupogimeno.com

[12] "AEMET OpenData," Jun. 2015, in Spanish. Accessed: 17 Oct. 2017. [Online]. Available: https://opendata.aemet.es

[13] "Ayuntamiento de Castellón. Planetario." May 1995, in Spanish. Accessed: 17 Oct. 2017. [Online]. Available: http://www.castello.es/archivos/598/img/

[14] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American statistical association*, vol. 58, no. 302, pp. 415–434, 1963.

[15] L. Breiman and R. Ihaka, *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California, 1984.

[16] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, ser. Springer Series in Statistics. Springer New York, 2009. [Online]. Available: https://books.google.es/books?id=tVIjmNS3Ob8C

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley, New York, 1973.

[19] D. J. Hand, "Discrimination and classification," *Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley, 1981*, 1981.

[20] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2004, vol. 544.

[21] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 2007.

[22] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine learning, neural and statistical classification*. Citeseer, 1994.

[23] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied statistics*, pp. 119–127, 1980.

[24] M. R. Segal, "Regression trees for censored data," *Biometrics*, pp. 35–47, 1988.

[25] H. Kim and W.-Y. Loh, "Classification trees with unbiased multiway splits," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 589–604, 2001.

[26] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[27] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data mining and knowledge discovery*, vol. 2, no. 4, pp. 345–389, 1998.

[28] H. Strasser and C. Weber, *On the asymptotic theory of permutation statistics*. SFB Adaptive Information Systems and Modelling in Economics and Management Science, WU Vienna University of Economics and Business, 1999.

[29] D. D. Jensen and P. R. Cohen, "Multiple comparisons in induction algorithms," *Machine Learning*, vol. 38, no. 3, pp. 309–338, 2000.

[30] O. J. Dunn, "Estimation of the means of dependent variables," *The Annals of Mathematical Statistics*, pp. 1095–1111, 1958.

[31] ——, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[32] P. H. Westfall and S. S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993, vol. 279.

[33] S. M. O'Brien, "Cutpoint selection for categorizing a continuous predictor," *Biometrics*, vol. 60, no. 2, pp. 504–509, 2004.

[34] T. Hothorn, K. Hornik, M. A. Van De Wiel, and A. Zeileis, "A lego system for conditional inference," *The American Statistician*, vol. 60, no. 3, pp. 257–263, 2006.

[35] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[36] A. Afifi and R. Elashoff, "Missing observations in multivariate statistics i. review of the literature," *Journal of the American Statistical Association*, vol. 61, no. 315, pp. 595–604, 1966.

[37] Y. Haitovsky, "Missing data in regression analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 67–82, 1968.

[38] M. Huisman, "Imputation of missing item responses: Some simple techniques," *Quality and Quantity*, vol. 34, no. 4, pp. 331–351, 2000.

[39] J.-O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," *Sociological Methods & Research*, vol. 6, no. 2, pp. 215–240, 1977.

[40] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[41] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[42] P. Hall, *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[43] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[44] ——, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[46] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[47] ——, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[48] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[49] M. R. Segal, "Machine learning benchmarks and random forest regression," *Center for Bioinformatics & Molecular Biostatistics*, 2004.

[50] E. Kleinberg *et al.*, "An overtraining-resistant stochastic modeling method for pattern recognition," *The annals of statistics*, vol. 24, no. 6, pp. 2319–2349, 1996.

[51] E. M. Kleinberg, "On the algorithmic implementation of stochastic discrimination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 473–490, 2000.

[52] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 128, 2010.

[53] P. Wei, Z. Lu, and J. Song, "Variable importance analysis: a comprehensive review," *Reliability Engineering & System Safety*, vol. 142, pp. 399–432, 2015.

[54] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, Jan 2007. [Online]. Available: https://doi.org/10.1186/1471-2105-8-25

[55] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010. [Online]. Available: +http://dx.doi.org/10.1093/bioinformatics/btq134

[56] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[57] A.-L. Boulesteix, S. Janitza, A. Hapfelmeier, K. Van Steen, and C. Strobl, "Letter to the editor: On the term 'interaction'and related phrases in the literature on random forests," *Briefings in bioinformatics*, vol. 16, no. 2, pp. 338–345, 2014.

[58] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC bioinformatics*, vol. 11, no. 1, p. 110, 2010.

[59] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random survival forests for high-dimensional data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 1, pp. 115–132, 2011.

[60] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 205–217, 2010.

[61] A. Pierola, I. Epifanio, and S. Alemany, "An ensemble of ordered logistic regression and random forest for child garment size matching," *Computers & Industrial Engineering*, vol. 101, pp. 455–465, 2016.

[62] I. Epifanio, "Intervention in prediction measure: a new approach to assessing variable importance for random forests," *BMC bioinformatics*, vol. 18, no. 1, p. 230, 2017.

[63] L. Breiman, "Manual on settings up, using and understanding random forest," *V4. 0, University of California Berkeley, Statistics Department, Berkeley*, 2003.

[64] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[65] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[66] H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008. [Online]. Available: http://arXiv.org/abs/0811.1645v1

[67] M. Segal and Y. Xiao, "Multivariate random forests," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 80–87, 2011.

[68] N. Patel and S. Upadhyay, "Study of various decision tree pruning methods with their empirical comparison in weka," *International journal of computer applications*, vol. 60, no. 12, 2012.

[69] C. Strobl and A. Zeileis, "Danger: High power! ? exploring the statistical properties of a test for random forest variable importance," 2008. [Online]. Available: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-2111-8

[70] K. V. Mardia, J. T. Kent, and J. M. Bibby, "Multivariate analysis (probability and mathematical statistics)," 1980.

[71] P. Diaconis, S. Goel, and S. Holmes, "Horseshoes in multidimensional scaling and local kernel methods," *The Annals of Applied Statistics*, pp. 777–807, 2008.

[72] E. Friedrich and D. Kretzinger, "Vulnerability of wastewater infrastructure of coastal cities to sea level rise: A south african case study," *Water SA*, vol. 38, no. 5, pp. 755–764, 2012.

[73] J. O. Ramsay, *Functional data analysis.* Wiley Online Library, 2006.

[74] F. Ferraty and P. Vieu, *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media, 2006.

[75] Y. A. Brüning, "Investigation of the relationship between sea level fluctuations and the electrical conductivity of wastewater," 2016.